Dunn 지수를 이용한 최적 강수지역 군집수 분석

The Analysis of Optimal Cluster Number of Precipitation Region with Dunn Index

엄명진*, 정창삼**, 남우성***, 정영훈****, 허준행*****

Myoung-Jin Um, Chang-Sam Jeong, Woo Sung Nam, Young Hun Jung, Jun-Haeng Heo

요 지

강수는 지역에 따라 발생양상이 매우 다른 자연현상 중 하나이다. 이러한 강수를 효과적으로 분석하여 확률강수량을 산정하기위해서 수문학에서는 다양한 방법이 시도되어 왔다. 우리나라에서는 지점빈도해석을 통한 확률강수량을 주로 사용해왔으나 최근 들어 Hosking and Wallis(1997)가 제안한 지역빈도해석을 활용을 적극 도모 하고 있는 중이다. 이러한 지역빈도해석 기법은 지점빈도해석 기법에 비하여 한정된 강수자료를 활용하는 측면 등 여러 가지 장점을 가진 확률강수량 산정방법이다. 그러나 이 기법을 적용하여 확률강수량을 산정하기 위해서는 강수의 지역구분을 먼저 수행하여야 한다.

강수지역의 구분을 위해서는 여러 가지 기법이 존재하나 최근에는 Cluster 기법 중 K-means 방법이나 Fuzzy c-means 방법 등을 주로 적용하여 지역구분을 수행하고 있다. 그러나 K-means 방법이나 Fuzzy c-means 방법 등은 산정 방법내에서 최적 군집수를 결정할 수 있는 알고리즘이 없기 때문에 임의적으로 최적 군집수를 결정하여야 한다. 본 연구에서는 이러한 단점을 극복하기 위하여 Cluster 평가지수 중 하나인 Dunn 지수를 이용하여 최적 군집수를 제시하고자 한다.

본 연구에서 강수지역을 구분하기 위하여 적용한 인자는 월 평균 강수량, 연 평균 강수량, 월 최대 강수량, 경도, 위도, 고도 등이며, 이를 K-means, PAM 및 친근도 전파 기법을 통하여 강수지역을 구분하였다. 적정 군집수를 임의적으로 증가시켜 가면서 Dunn 지수를 산정하였다. 산정된 결과를 통하여 최적 군집수를 결정하였다.

핵심용어 : 지역빈도해석, Cluster 기법, Dunn 지수

1. 서론

지점 빈도해석은 자료수가 부족할 때 확률 수문량을 산정하는데 있어 정확도가 떨어지는 단점을 갖고 있다. 이러한 문제점을 보완할 수 있는 방법인 지역빈도 해석 방법은 우리나라와 같이 자료수가 부족하거나 미 계측 지점에서에서 효율적이고 안정적으로 확률 수문량을 산정할 수 있다(이동진, 허준행, 2001; 허준행 등, 2004; Cunnane, 1989).

지수홍수법 등을 이용한 지역빈도해석을 적용하기 위해 정해진 지역의 모든 지점들이 동일한 수문학적 특성을 가진다고 간주할 경우 서로 다른 지점들의 공간 특성으로 인하여 지점 간의 이질성이 증가하여 정확도가 떨어질 수 있는 단점을 가지고 있다(김경덕, 허준행, 2007; Gabriele and Arnell, 1991).

지역빈도해석 단계 중 동질 지역의 구분은 중요한 단계라 할 수 있다. 그러나 동질 지역 구분을 위한 정

* 정회원・연세대학교 토목공학과 박사・E-mail : movie21@yonsei.ac.kr
** 정회원・인덕대학교 토목환경공학과 교수・E-mail : csjeong@induk.ac.kr
*** 정회원・연세대학교 토목공학과 박사과정・E-mail : nws77@yonsei.ac.kr
**** 정회원・연세대학교 토목공학과 박사과정・E-mail : yhjung2000@yonsei.ac.kr
***** 정회원・연세대학교 토목공학과 교수・E-mail : jhheo@yonsei.ac.kr

해진 기준은 없기 때문에 이를 위한 다양한 기법과 변수들이 활용되어 왔다. 일례로, Mallants and Feyen (1990)은 일강우량 자료만을 활용하여 지역을 구분하였고, Guttman (1993)은 7개의 지형 및 기후와 관련된 변수를 바탕으로 지역을 구분하였다. Zhang and Hall (2004)은 몇 가지 군집해석 기법을 활용해서 지역을 구분했고, Dinpashoh et al. (2004)은 주성분 분석, Procrustes Analysis, 요인분석을 활용하여 지역을 구분함으로 동질 지역 구분의 효율을 향상에 대해 연구하였다(남우성 등, 2008).

2. 강수지역구분

2.1 요인분석

요인분석은 변수가 많은 경우에 군집해석의 효율성 저하 문제를 해결하기 위해 주로 사용되는 방법이다. 요인분석은 기존 변수들의 상관성을 이용하여 요인(factor)이라 불리는 변수 내의 공통적인 새로운 변수를 도출하여 이 요인들이 가지고 있는 특성으로 전체 자료의 특성을 최대한 설명하는 기법이다.

본 연구에서는 다양한 변수에 대하여 통계 패키지인 SPSS을 활용하여 요인을 선정하였으며, 요인추출은 최우도법, 요인회전은 기후자료에 적합하다고 알려진 varimax rotation (Overall and Klett, 1973; Puvaneswaran, 1990; White et al., 1991)을 적용하였다.

2.2 군집분석

군집 분석(Cluster analysis)은 집락분석(집단을 규명하는 방법)이라 불리기도 하며, 개체들이 지니고 있는 다양한 속성의 유사성을 동질적인 집단으로 군집화하는 방법을 말한다. 이는 개체들이 일정한 속성에 따라 몇 개의 군집으로 나누어 각 집단 간의 상관성을 이해하고 효율적으로 이용할 수 있게 하는 것을 말한다. 군집분석 기본원리는 분석하고자 하는 여러 특성들을 유사성(Similarity) 거리(Distance)로 환산하고 거리가 상대적으로 가까운 개체들을 동질적으로 군집화 하는 것이다(박상우, 2003). 이러한 군집분석은 어느 한 기법이 우수하다고 단정하기가 매우 어려우며 자료의 상황에 맞게 적절한 군집방법과 군집개수를 산정해야 한다.

가. k-means 군집분석

통계학이나 데이터마이닝에서 k-means 군집은 가장 가까운 평균을 가진 군집에 속해 있는 각각의 관측치안에 n개의 관측치들을 k개의 군집으로 분할하려는 목적을 가진 군집분석 방법이다. 이러한 k-means 알고리즘(McQueen, 1967)은 주어진 자료에 안에서 군집을 확인할 때 흔히 사용되는 방법이다.

각각의 관측치들이 d차원의 실제 벡터인 관측치들($x_1, x_2, \cdots, x_n$)이 주어졌을 경우, k-means 군집은 n개의 관측치들을 k개의 집합($k \leq n$)으로 분할하는데 목적이 있다. 또한 군집간 제곱합을 식 (1)과 같이 최소하여야 한다.

$$\arg min_S \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \qquad (1)$$

여기서, μ_i는 S_i상의 자료들의 평균이다.

나. Partitioning Around Medoids(PAM) 알고리즘

PAM 알고리즘은 k-medoids 알고리즘으로 불리우며 k-means방법과 유사하다. 다만, k-means 기법은 극도로 큰 값(혹은 작은 값)이 자료의 분포를 사실상 왜곡할 수 있기 때문에 이상치에 민감하다. 이를 위해 군집에서 객체들의 평균값을 취하는 대신에 군집에서 가장 중심에 위치한 객체인 medoid를 사용할 수 있다. k-medoids 군집화 알고리즘의 기초적인 방법은 각 군집에서 대표 객체(medoids)를 임의로 찾음으로써 n개의 객체 중에서 k개의 군집을 찾는 것이다. 남은 각각의 객체는 가장 비슷한 medoids에 군집된다.

이러한 산정절차를 정리하면 다음과 같다.

Step 1. n개의 객체 중 대표 객체(medoids)를 k개 지정한다. (n>k)

Step 2. k개의 medoids를 지정 후, 나머지 객체를 유사성이 가장 높은 medoid에 배속한다. 여기서 유사성은 거리측도를 활용한다.

Step 3. medoid가 아닌 다른 객체를 임의로 지정한다.

Step 4. 본래의 medoid와 임의의 medoid간의 총 거리측도 합을 계산한다.

Step 5. 만약 총 거리측도 합이 음수인 경우, 임의의 medoid를 기존의 medoid와 교체한다.

Step 6. 변화가 없을 때까지 Step2~Step5를 반복한다

다. 친근도 전파(Affinity Propagation) 군집

Frey and Dueck(2007)은 친근도 전파에 기반을 둔 군집 분석 알고리즘을 제안하였다. 자료들간의 유사도를 반영하는 메시지를 자료들끼리 주고받음으로써 각 자료마도 자신을 대표하는 점(exempler)가 결정되고, 동일 대표점을 갖는 자료들기리 하나의 군집으로 지정되는 것이다. 이때 메시지는 각 자료와 지정된 대표 점들 간의 유사도를 극대화시킬수 있도록 가정된다. 친근도 전파는 각 자료에 대한 대표점을 선정하기 위한 두 종류의 척도를 나타내는 책임도(responsibility)와 유효도(availability)라는 메시지가 있다. 두 종류의 메시지는 모두 I번째와 j번째의 자료의 유사도를 나타내는 입력정보 s(i,j)에 의해 결정된다(이수찬 등, 2008).

2.3 유효성측도

원 자료로서 군집 결과를 분석하여 자료 본래의 정보를 사용하여 군집화가 잘 되었는지를 판단하는 측도를 일반적으로 내부유효성 측도라 하며, Dunn Index(Dunn, 1974) 및 Silhouette Width(Rousseeuw, 1987) 등이 있다. 본 연구에서는 Dunn Index를 통하여 군집유효성을 판단하려 한다.

Dunn 지수는 같은 군집에 속해 있는 두 개체간의 가장 큰 거리에 대한 서로 다른 군집에 속해 있는 두 개체간의 가장 작은 거리 비(ratio)를 나타내면 식 (2)와 같이 산정할 수 있다.

$$V_D = \min_{1 \le i \le k} \left\{ \min_{1 \le j \le k, j \ne i} \left[\frac{\delta(C_i, C_j)}{\max_{1 \le k \le K} \Delta(C_k)} \right] \right\} \quad (2)$$

여기서 $\delta(C_i, C_j)$은 군집 C_i와 C_j의 거리를 나타내며, $\Delta(C_k)$는 C_k에 대한 군집간 거리를 말한다. 같은 군집에 속해 있는 두 개체간의 거리가 작을수록 다른 군집에 속해있는 두 개체간의 거리가 클수록 Dunn 지수는 커지므로 Dunn 지수가 클수록 군집이 잘 이루어졌다고 판단할 수 있다.

3. 대상지역 및 지역구분

본 연구에서는 전국의 기상청 산하 68개 강우 관측 지점의 강우자료를 바탕으로 분석을 수행하였다. 68개 지점은 우리나라 전역에 분포되어 있으며, 자료 기간은 10~88년에 이른다. 동질 지역 구분을 위해 Table 1과 같이 강우에 영향을 미치는 변수들을 68개 강우 관측 지점에 대해 수집하였다.

Table 1. Climatological and Geographical Candidate Variables Related with Precipitation

Variable	Description
LAT	LATITUDE
LONG	LONGITUDE
ALT	ALTITUDE
MAP	Mean Annual Precipitation
APM_i, i = 1, 2,…,12	Average Precipitation in a Month
MDP_i, i = 1, 2, … ,12	Average Maximum Daily Precipitation in a month
AMaxMDP	Average Maximum of Maximum Daily Precipitation in a month

수집된 자료들에 대하여 요인분석을 실시하여 5개의 요인을 산정하였다. 요인분석 결과 Kaiser-Meyer-Olkin 측도는 0.52로 0.5보다 크게 나타났으며 5개 요인의 누적 설명 총분산은 89.14%이다. 이렇게 산정된 5개 요인을 대상으로 군집분석을 실시하였다. 군집분석 방법으로는 k-means, PAM, 친근도 전파 방법을 적용하였다. 군집수는 3개에서 9개까지를 가정하여 적용하였다. 각각 분할된 군집들에 대하여 유효성측도인 Dunn 지수를 산정하여 Fig. 1에 도시하였다. 도시 결과 세 가지 분석 방법이 모두 군집수가 5개 일 경우 Dunn 지수가 크게 나타났다. 다만 k-means 방법의 경우 군집수 3개와 5개 일 경우 그 그 값이 각각 3.716, 3.752로 유사하게 산정되었다.

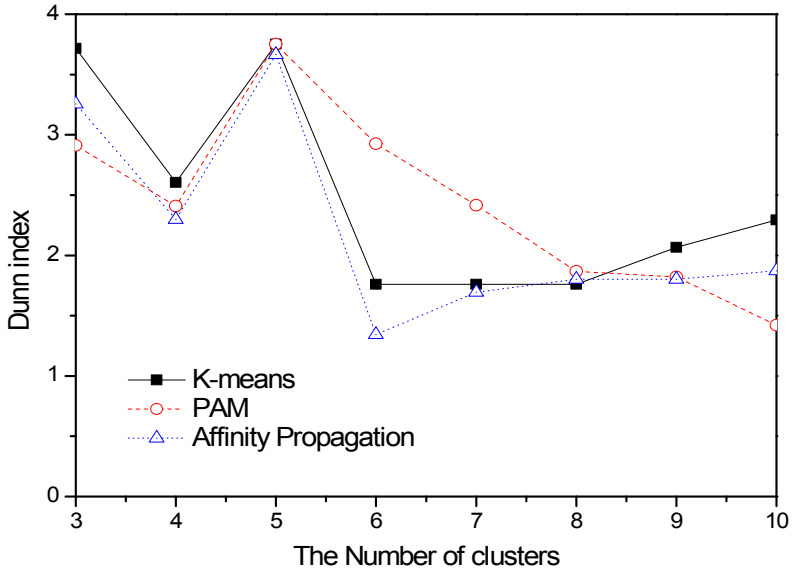

Fig. 1. Dunn Index with three methods of cluster technique

4. 결론

본 연구에서는 68개 강우 관측소의 자료를 기반으로 강수지역을 구분하였다. 강수지역 구분을 위하여 관측지점별 지형자료 및 강우관련 자료를 수집하였으며, 요인분석을 통하여 관련 항목을 29개에서 5개 요인으로 최소화하면서 자료의 중복효과를 제거하여 분석의 효율성을 높였다. 지점별로 산정된 5개 요인에 대하여 3개의 군집분석을 군집수를 3개에서 9개로 변화시켜가면서 Dunn 지수의 변화양상을 분석하였다. 분석결과 세가지 군집방법 모두 군집수 5개일 경우 Dunn 지수가 높게 산정되었다. 그러므로 분석대상지역은 군집수가 5개가 적합한 것으로 판단된다. 따라서 본 연구에서는 요인분석과 군집유효성 측도인 Dunn 지수를 통하여 각 군집기법별로 최적 군집수를 선정하기 위한 자료를 제공함으로써 강수 지역구분을 용이하게 수행할 수 있도록 하였다. 하지만 본 연구결과는 보다 정확한 검증을 위하여 추후 지역빈도해석 알고리즘에 포함된 동질성 검사 및 해석결과에 따른 제곱근편차 등을 계산하여 재검증을 수행하여야 할 것으로 판단된다.

감 사 의 글

본 연구는 국토해양부 지역기술혁신사업의 연구비지원(09-지역기술혁신 B-01-02, 수충부 및 토석류 방재기술연구사업)에 의해 추진되었습니다

참 고 문 헌

1. 김경덕, 허준행 (2007). 모의실험을 통한 지수홍수법의 수행능력 해석 연구, 대한토목학회논문집, 제27권 제1호, pp.9-20.
2. 남우성, 김태순, 신주영, 허준행 (2008). 다변량 분석 기법을 활용한 강우 지역빈도해석, 한국수자원학회 논문집, 제41권 제5호, pp. 517-525.
3. 박상우, 전병호, 장석환 (2003). 다변량 분석기법에 의한 지점강우의 권역화 연구, 한국수자원학회논문집, 제36권 제5호, pp.879-892
4. 이동진, 허준행 (2001). L-모멘트법을 이용한 한강유역 일강우량자료의 지역빈도해석, 한국수자원학회논문집, 제34권 제2호, pp119-130.
5. 이수찬, 박상현, 윤일동, 이상욱 (2008). 검색과 분류를 위한 친근도 전파 기반 3차원모델의 특징적 시점 추출 기법, 한국방송공학회 논문집, 제13권 제6호, pp. 828-837.
6. 허준행, 이영석, 남우성, 김경덕 (2004). “한강유역에 대한 강우지역빈도해석의 적용성 연구”, 한국수자원학회 학술발표회 논문집.
7. Cunnane, C. (1989). Statistical distributions for flood frequency analysis, Hydrol. Rep. No. 33, WMO Publ. No. 718, Geneva.
8. Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S., Mirnia, M. (2004). Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods, Journal of Hydrology, Vol. 297, pp. 109-123.
9. Dunn, J.C. (1974) Well-separated clusters and optimal fuzzy partitions, Journal of Cybernetics, Vol 4, pp. 95-104.
10. Frey, B. J., Dueck, D. (2007) Clustering by passing messages between data points. Science, Vol. 315, pp. 972-976.
11. Gabriele, S.G., Arnell, N. (1991). A Hierarchical Approach to Regional Flood Frequency Analysis, Water Resources Research, Vol. 27, No. 6, pp. 1281-1289.
12. Guttman, N.B. (1993). The use of L-Moments in the determination of regional precipitation climates, Journal of Climatology, Vol. 6, pp. 2309-2325.
13. Hosking, J.R.M. and Wallis, J.R. (1997). Regional frequency analysis: an approach based on L-moments, Cambridge University Press.
14. Mallants, D., Feyen, J. (1990). Defining homogeneous precipitation regions by means of principal component analysis, Journal of Applied Meteorology, Vol. 29, pp. 892-901.
15. McQueen, J.B. (1967) Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, Vol. 1, pp. 281-297.
16. Overall, J.E. and Klett, C.J. (1972). Applied multivariate analysis, McGraw-Hill, New York.
17. Puvaneswaran, M. (1990). Climatic classfication for Queensland using multivariate statistical techniques. Int. J. Climatol. Vol. 10, pp. 591-608.
18. Rousseeuw, P.J. (1987) Silhouettes: Graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Vol. 20, pp. 53-65.
19. White, D., Richman, M., and Yanel, B. (1991). Climate regionalization and rotation of principal components. Int. J. Climatorl. Vol. 11, pp. 1-25.
20. Zhang Jingyi, M.J. Hall (2004). Regional flood frequency analysis for the Gan-Ming River basin in China, Journal of Hydrology, Vol. 296, pp. 98-117.